4-Ч

УДК 004.62:004.023

***О.Р. Чертов, Д.Ю. Тавров***
Національний технічний університет України «Київський політехнічний інститут», м. Київ
Україна, 03056, м. Київ, пр-т Перемоги, 37

# Меметичний алгоритм для модифікації мікрофайлу з мінімізацією спотворень у процесі забезпечення групової анонімності

***O.R. Chertov, D.Y. Tavrov***
*National Technical University of Ukraine «Kyiv Polytechnic Institute», c. Kyiv*
*Ukraine, 03056, c. Kyiv, Peremohy ave., 37*

# *Memetic Algorithm for Microfile Modification With Minimizing Distortion While Providing Group Anonymity*

***О.Р. Чертов, Д.Ю. Тавров***
Национальный технический университет Украины «Киевский политехнический институт», г. Киев
Украина, 03056, г. Киев, пр-т Победы, 37

# Меметический алгоритм для модификации микрофайла с минимизацией искажений в процессе обеспечения групповой анонимности

У статті розглянуто задачу модифікації мікрофайлу зі статистичними даними для забезпечення анонімності даних про певні групи респондентів. Виконано огляд існуючих евристичних методів розв’язання цієї задачі та запропоновано новий меметичний алгоритм її розв’язання. Проведено порівняльний аналіз евристичних методів та меметичного алгоритму на основі прикладу з реальними даними.
**Ключові слова:** мікрофайл, групова анонімність, еволюційний алгоритм, меметичний алгоритм.

In the article, the task of modifying the microfile of statistical data for providing group anonymity of certain respondent group data is discussed. Existent heuristic methods of solving this task are described, and a novel memetic algorithm for solving the task is proposed. Heuristic methods and memetic algorithm are compared in performance on a real data based example.
**Key words:** microfile, group anonymity, evolutionary algorithm, memetic algorithm.

В статье рассмотрена задача модификации микрофайла со статистическими данными для обеспечения анонимности данных об определенных группах респондентов. Осуществлен обзор существующих эвристических методов решения этой задачи и предложен новый меметический алгоритм ее решения. Проведен сравнительный анализ эвристических методов и меметического алгоритма на основе примера с реальными данными.
**Ключевые слова:** микрофайл, групповая анонимность, эволюционный алгоритм, меметический алгоритм.

## Вступ

Сучасні інформаційні технології дозволяють аналізувати великі обсяги первинних неагрегованих даних, на базі яких дослідники можуть висувати гіпотези та перевіряти їхню справедливість. Проект IPUMS-International [1], у рамках якого на

момент написання даної роботи зібрано близько 480 млн. записів про респондентів 211 переписів населення в 68 країнах, є одним із найпоказовіших прикладів доступності первинних даних для аналізу.

Публікація первинних даних підвищує ризик розкриття чутливої інформації про окремих осіб або групи осіб. Перед публікацією даних потрібно забезпечувати їхню анонімність. Під анонімністю розумітимемо неідентифіковність суб'єкта в рамках певної множини суб'єктів [2]. Захист інформації про особу передбачає виконання *індивідуальної* анонімізації даних, захист внутрішніх закономірностей, особливостей розподілу даних – *групової* анонімізації.

Під час групової анонімізації даних потрібно забезпечувати приховання чутливих особливостей їх розподілу і водночас зберігати достатній рівень їхньої корисності [3]. Існуючі методи групової анонімізації розв'язують цю задачу в два етапи. На першому етапі відбувається видозміна розподілу даних з метою приховання чутливих особливостей, на другому – фізична модифікація мікрофайлу для встановлення відповідності між даними та їх видозміненим розподілом. Мінімізація спотворень мікрофайлу на другому етапі – складна оптимізаційна задача, розв'язання якої передбачає застосування евристичних стратегій [4, с. 258].

Еволюційні алгоритми являють собою евристичні методи цілеспрямованого випадкового пошуку, що моделюють процес еволюції за допомогою природного відбору [5, с. 15].

Якщо популяція індивідів існує в середовищі з обмеженими ресурсами, боротьба за доступ до цих ресурсів спричиняє природний відбір, який, у свою чергу, веде до збільшення середньої пристосованості індивіда в популяції до середовища. Аналогами індивідів в еволюційному алгоритмі є конкретні розв'язки деякої задачі, якість яких називають пристосованістю.

Основні етапи еволюційного алгоритму передбачають відбір так званих батьківських індивідів, їхнє схрещування між собою та мутацію нащадків.

Серед батьків та їхніх нащадків на основі значень функції пристосованості відбираються ті індивіди, які буде поміщено в популяцію нового покоління.

Вказані етапи мають стохастичну природу, що забезпечує ефективний пошук у просторі можливих розв'язків, але в той же час уповільнює збіжність алгоритму в околі оптимуму.

Меметичні алгоритми, уперше запропоновані в [6], моделюють явище культурної еволюції, яка відбувається на рівні мемів [7, с. 192]. Із математичного погляду, меметичні алгоритми часто реалізують як еволюційні, у яких на одному або декількох основних етапах вводиться операція локального пошуку. Застосування локального пошуку дозволяє [5, с. 174] підвищити ефективність еволюційного алгоритму для розв'язання окремих класів задач за рахунок урахування їхніх особливостей, а також підвищити збіжність алгоритму в околі оптимуму.

**Метою даної роботи** є розробка меметичного алгоритму для мінімізації спотворень у мікрофайлі в процесі його групової анонімізації, а також порівняння його ефективності з описаними в літературі евристичними стратегіями на базі прикладу з реальними даними.

# Схема забезпечення групової анонімності

Під *мікрофайлом* $M$ розумітимемо дані про респондентів, зібрані в файл записів $r_i$, $i = \overline{1,m}$, які містять значення атрибутів $w_j$, $j = \overline{1,\eta}$. Значення $j$-го атрибуту запису $r_i$ позначатимемо $z_{ij}$. Вважатимемо, що мікрофайл є знеособленим, тобто серед його атрибутів немає *ідентифікаторів*, що однозначно визначають респондента.

Позначимо через $w_j$ множину значень атрибуту $w_j$. *Сутнісна множина* $V = \{V_1, V_2, \ldots, V_{l_v}\}$ – це підмножина декартового добутку значень *сутнісних атрибутів* $w_{v_1} \text{Ч} w_{v_2} \text{Ч}...\text{Ч} w_{v_t}$, $v_j \in Z$ "$j = \overline{1,t}$. Кожен елемент $V_k \in V$, $k = \overline{1, l_v}$, $l_v \le |w_{v_1}| \text{Ч} |w_{v_2}| \text{Ч}...\text{Ч} |w_{v_t}|$, де $|\cdot|$ позначає потужність множини, називають *сутнісною комбінацією значень*, яка складається з *сутнісних значень*. Сутнісні множини дозволяють описати підмножини респондентів, розподіл яких потрібно приховати.

*Параметризуюча множина* $P = \{P_1, P_2, \ldots, P_{l_p}\}$ – це підмножина $w_p$, $p \neq v_j$ $\forall j = \overline{1,t}$. Атрибут $w_p$ називають *параметризуючим*, а елементи $P_k \in P$, $k = \overline{1, l_p}$, $l_p \le |w_p|$ — *параметризуючими значеннями.* Вони дозволяють упорядкувати дані мікрофайлу. Параметризуючі значення розбивають $M$ на *підмікрофайли* $M_1,\ldots,M_{l_p}$.

*Група* $G(V,P)$ – множина [8] сутнісних комбінацій значень $V$ та параметризуючих значень $P$, визначених для певної задачі групової анонімності.

Під *задачею забезпечення групової анонімності* розуміють модифікацію початкового набору даних для кожної групи $G_i(V_i,P)$, $i = \overline{1,k}$ таким чином, щоб чутливі (для розв'язання даної задачі) властивості даних було захищено. Початковий набір даних $M$ потрібно змінювати окремо для кожної групи. Загальна схема за безпечення групової анонімності передбачає [9] виконання наступних кроків:

1. Підготовка знеособленого мікрофайлу $M$.
2. Визначення груп $G_i(V_i,P)$, $i = \overline{1,k}$, що представляють категорії респондентів, розподіл яких потрібно приховати.
3. Для кожного *i* від 1 до *k*:

а) вибір *цільового представлення* $\Omega(M,G_i)$, яке визначає набір даних довільної структури, що представляє особливості групи в спосіб, зручний для модифікації;

б) виконання відображення даних за допомогою *цільового відображення* $\Upsilon$: $M \to \Omega_i(M, G_i)$;

в) одержання *модифікованого цільового представлення* за допомогою *модифікуючого функціоналу* $\Xi$: $\Omega_i(M,G_i) \to \Omega^*_i(M,G_i)$ таким чином, щоб чутливі особливості набору даних стали прихованими;

г) одержання модифікованого мікрофайлу шляхом застосування *оберненого цільового відображення* $\Upsilon^{-1}$: $\Omega^*_i(M,G_i) \to M^*$.

4. Підготовка модифікованого мікрофайлу $M^*$ до публікації.

Розгляньмо одне часто використовуване цільове представлення – *кількісний сигнал* $q = (q_1, q_2, \ldots, q_{l_p})$, який складається з кількостей респондентів з однаковими сутнісними комбінаціями значень та параметризуючими значеннями. Для спрощення вважатимемо, що кожному підмікрофайлу $M_i$, $i = \overline{1,l_p}$ відповідає одне значення кількісного сигналу. Різницю $\delta = q - q*$ між кількісним та модифікованим кількісними сигналами називають *різницевим сигналом*.

*Валентністю підмікрофайлу* $M_i$ називають [4] значення $\delta_i$. Позначимо через $I^+ = \{i \mid \delta_i > 0\}$ – множину індексів підмікрофайлів із додатними валентностями, а через $I^- = \{i \mid \delta_i < 0\}$ – множину індексів підмікрофайлів із від'ємними валентностями.

# Мінімізація спотворень мікрофайлу під час його модифікації

*Визначальні атрибути* – такі, розподіл яких параметризуючими значеннями становить інтерес для дослідників [10]. *Визначальна метрика* – функція, яка двом записам мікрофайлу $r_1$ та $r_2$ зіставляє число, що обчислюється за формулою

$$\mathrm{InfM}(r_1, r_2) = \sum_{l=1}^{n_{ord}} \omega_l \left( \frac{r_1(I_l) - r_2(I_l)}{r_1(I_l) + r_2(I_l)} \right)^2 + \sum_{k}^{n_{nom}} \gamma_k \chi^2 \left( r_1(J_k), r_2(J_k) \right), \quad (1)$$

де $I_l$ – $l$-ий визначальний атрибут порядкового типу, загальна кількість котрих дорівнює $n_{ord}$, $J_k$ – $k$-ий визначальний атрибут номінального типу, загальна кількість котрих дорівнює $n_{nom}$, $r(\cdot)$ – оператор повернення значення відповідного атрибуту запису $r$, $\omega_l$ та $\gamma_k$ – невід'ємні вагові коефіцієнти, $\chi(v_1, v_2)$ – функція, що дорівнює $\chi_1$, якщо $v_1$ та $v_2$ належать одній категорії, та $\chi_2$ – в іншому випадку.

Мінімізації спотворень даних мікрофайлу можна досягти обміном параметризуючих значень між записами, близькими в розумінні (1). Зрозуміло, що на практиці неможливо [4] перебрати всі пари записів $\langle r_1, r_2 \rangle$ з метою вибору тих, обмін параметризуючих значень котрих мінімізуватиме сумарне значення (1). Потрібно використовувати евристичні стратегії [4] підбору пар записів $\langle r_1, r_2 \rangle$, котрі дають результати, прийнятні з погляду обчислювальної складності та близькості до мінімуму (1).

У літературі запропоновано [4] низку таких стратегій для кількісного сигналу. Загальна схема всіх стратегій передбачає виконання наступних кроків:

1. Підрахувати різницевий сигнал $\delta$.
2. Вибрати підмікрофайл $M_i$, $i \in I^+$.
3. Вибрати запис $r_k \in M_i$, значення сутнісних атрибутів якого дорівнюють сутнісній комбінації значень.
4. Вибрати підмікрофайл $M_j$, $j \in I^-$.
5. Вибрати запис $r_l \in \mathrm{M}_j$, значення сутнісних атрибутів якого не дорівнюють жодній сутнісній комбінації значень, та який є найближчим у розумінні (1) до $r_k$.
6. Обміняти параметризуючі значення між $r_k$ та $r_l$. Зменшити на одиницю за абсолютним значенням $\delta_i$ та $\delta_j$.
7. Якщо $\sum_p \left| \delta_p \right| = 0$, вийти. Інакше, перейти на крок 2.

Усі стратегії відрізняються реалізацією кроків 2, 3 та 4. Можливі реалізації кроків 2 та 4 представлено в табл. 1 (номер стратегії 10 умисно пропущено). На кроці 3 для стратегій 1 – 9 запис $r_k$ вибирають випадковим чином, для стратегій 11 – 19 – перебирають усі можливі записи та вибирають той, що на кроці 5 виявиться найближчим у розумінні (1).

Таблиця 1 – Стислий опис евристичних стратегій модифікації мікрофайлу

| Номер стратегії | Вибір підмікрофайлу на кроці 2 | Вибір підмікрофайлу на кроці 4 |
|---|---|---|
| 1, 11 | із найменшим індексом | із найменшим індексом |
| 2, 12 | із найбільшою валентністю | із найменшою валентністю |
| 3, 13 | із найменшою валентністю | із найбільшою валентністю |
| 4, 14 | із найменшим індексом | із найбільшою кількістю записів |
| 5, 15 | із найбільшою валентністю | |
| 6, 16 | із найменшою валентністю | |
| 7, 17 | із найменшим індексом | такий, що містить запис $r_l$, найближчий до обраного на кроці 3 в розумінні (1) |
| 8, 18 | із найбільшою валентністю | |
| 9, 19 | із найменшою валентністю | |

# Структура меметичного алгоритму для модифікації мікрофайлу

Загальна структура меметичного алгоритму для модифікації мікрофайлу $M$ передбачає виконання наступних кроків.

1. Випадковим чином згенерувати популяцію $P$ з $\mu$ індивідів та застосувати до кожного індивіда оператор локального пошуку $S$.
2. Обчислити значення функції пристосованості $f(x)$ для кожного індивіда $x \in P$.
3. Перевірити умову завершення. Якщо вона виконується, завершити, інакше – перейти на крок 4.
4. Відібрати $\lambda$ пар батьківських індивідів.
5. Застосувати оператор схрещування $R$ до кожної пари батьківських індивідів, отримуючи по два нащадки після кожного схрещування.
6. Застосувати оператор мутації $M$ до кожного з нащадків та помістити результуючих індивідів у множину $P'$.
7. Застосувати оператор локального пошуку $S$ до кожного $\mathbf{x} \in P'$.
8. Обчислити значення функції пристосованості $f(\mathbf{x})$ для кожного $\mathbf{x} \in P'$.
9. Відібрати $\mu$ найбільш пристосованих індивідів із $P \cup P'$ та помістити їх у $P$ замість поточних.
10. Перейти на крок 3.

Індивіди в популяції повинні представляти конкретні розв'язки задачі модифікації мікрофайлу. Доцільним видається обрати наступне *представлення* розв'язків задачі в рамках меметичного алгоритму. Кожен індивід є матрицею $U = \|u\|_{Q\times 4}$, у якій елементи першого стовпця – це індекси підмікрофайлів із додатними валентностями ($u_{i1} \in I^{+} \quad \forall i = \overline{1,Q}$), третього стовпця – індекси підмікрофайлів із від'ємними валентностями ($u_{i3} \in I^{-} \quad \forall i = \overline{1,Q}$). Елементи другого стовпця $u_{i2}$ виступають номерами записів із підмікрофайлу $M_{u_{i1}}$, четвертого $u_{i4}$ – із підмікрофайлу $M_{u_{i3}}$.

Таким чином, кожен рядок матриці $U$ визначає пару записів із різних підмікрофайлів, параметризуючі значення яких потрібно обміняти. При цьому перший та другий стовпці матриці $U$ визначають записи, значення сутнісних атрибутів котрих дорівнюють сутнісній комбінації значень, а третій та четвертий – записи, значення сутнісних атрибутів котрих не дорівнюють жодній сутнісній комбінації значень.

На елементи матриці $U$ накладаються певні структурні обмеження:

1. Індекс $i^+$ кожного підмікрофайлу з додатною валентністю $M_{i^+}$ повинен зустрічатися в першому стовпці матриці $U$ $\delta_{i^+}$ разів, тобто перший стовпець матриці $U$ є перестановкою над мультимножиною індексів $i^+ \in I^+$, кожен із яких узято $\delta_{i^+}$ разів.

2. Індекс $i^-$ кожного підмікрофайлу з від'ємною валентністю $M_{i^-}$ повинен зустрічатися в третьому стовпці матриці $U$ $\left(-\delta_{i^-}\right)$ разів, тобто третій стовпець матриці $U$ є перестановкою над мультимножиною індексів $i^- \in I^-$, кожен із яких узято $\left(-\delta_{i^-}\right)$ разів.

3. Кожна пара $\langle u_{i1}, u_{i2} \rangle$ або $\langle u_{i3}, u_{i4} \rangle$ $\forall i = \overline{1,Q}$ має зустрічатися в $U$ тільки 1 раз.

У якості *функції пристосованості* індивіда $f(U)$ доцільно взяти наступну:

$$f(U) = C_{\max} - \sum_{i=1}^{Q} InfM\left(M_{u_{i1}}\left(u_{i2}\right), M_{u_{i3}}\left(u_{i4}\right)\right), \tag{2}$$

де $M\left(i\right)$ – оператор повернення $i$-о запису мікрофайлу $M$, $C_{\max}$ – терм, що обчислюється за наступною формулою:

$$C_{\max} = \mathrm{InfM}_{\max} \cdot Q, \tag{3}$$

де $\mathrm{InfM}_{\max}$ – максимально можливе значення (1).

Функція пристосованості (2) – монотонно спадна функція сукупного значення (1), накопиченого в процесі переміщення $Q$ записів.

Терм $C_{\max}$ гарантує невід'ємність $f(U)$.

*Оператор схрещування* $R(U_{p1}, U_{p2})$ застосовується з високою ймовірністю $p_c$ до двох батьківських індивідів $U_{p1}$ та $U_{p2}$ та повертає двох індивідів-нащадків $U_{o1}$ та $U_{o2}$. Оператор не повинен порушувати структурні обмеження на $U$ № 1 та № 2.

*Оператор мутації* $M(U)$ застосовується до одного індивіда $U$ та повертає модифікованого індивіда $U'$. У силу особливостей представлення індивідів пропонується визначити оператор мутації як суперпозицію чотирьох операторів $M = M_4 \circ M_3 \circ M_2 \circ M_1$:

– $M_1$ – будь-який оператор, що з малою ймовірністю $p_{m_1}$ діє на перший стовпець матриці $U$ як на перестановку; переміщення елемента $u_{i1}$, $i \in \{1, 2, \ldots, Q\}$ з одного рядка в інший має наслідком аналогічне переміщення елемента $u_{i2}$;

– $M_2$ – будь-який оператор, що з малою ймовірністю $p_{m_2}$ діє на третій стовпець матриці $U$ як на перестановку; переміщення елемента $u_{i3}$, $i \in \{1, 2, \ldots, Q\}$ з одного рядка в інший має наслідком аналогічне переміщення елемента $u_{i4}$;

– $M_3$ – будь-який оператор, що з малою ймовірністю $p_{m_3}$ діє на другий стовпець матриці $U$ як на вектор номінальних цілочисельних значень; у процесі застосування оператора $M_3$ не повинно порушуватися структурне обмеження на $U$ № 3;

– $M_4$ – будь-який оператор, що з малою ймовірністю $p_{m_4}$ діє на четвертий стовпець матриці $U$ як на вектор номінальних цілочисельних значень; у процесі застосування оператора $M_4$ не повинно порушуватися структурне обмеження на $U$ № 3.

*Оператор локального пошуку* $S(U)$ застосовується до одного індивіда $U$ та повертає модифікованого індивіда $U'$. У даній роботі в якості такого оператора пропонується обрати оператор, який передбачає виконання наступних кроків.

1. Для кожного рядка $i$, $i = \overline{1,Q}$ матриці $U$ виконати кроки 2 – 4.
2. Згенерувати випадкове число $r \in [0,1]$.
3. Якщо $r \le p_{mem}$, записати в якості $u_{i4}$ індекс запису з підмікрофайлу $M_{u_{i3}}$, найближчого до запису $u_{i2}$ з підмікрофайлу $M_{u_{i1}}$ у розумінні (1). Якщо $r > p_{mem}$, записати в якості $u_{i2}$ індекс запису з підмікрофайлу $M_{u_{i1}}$, найближчого до запису $u_{i4}$ з підмікрофайлу $M_{u_{i3}}$ в розумінні (1).
4. Перейти на крок 2.

У процесі застосування $S$ не повинно порушуватися структурне обмеження на $U$ № 3. Параметр $p_{mem}$ має бути доволі великим, щоб забезпечити ефективність локального пошуку, але меншим 1, щоб упередити зупинку алгоритму в околі локального оптимуму.

Відбір батьківських індивідів та індивідів, які буде перенесено в популяцію наступного покоління, не залежить від представлення. Інші параметри меметичного алгоритму, наприклад, розмір популяції $\mu$, кількість нащадків у кожному поколінні $\lambda$, умову завершення роботи потрібно підбирати окремо для кожної задачі.

# Практичні результати

Для ілюстрації застосування меметичного алгоритму для модифікації мікрофайлу розглянемо задачу приховання територіального розподілу військових штату Массачусетс (США) на основі п’ятивідсоткової вибірки перепису населення США 2000 р. [11].

У якості параметризуючого атрибуту візьмемо атрибут «Place of Work Super-PUMA» («Місце роботи Super-PUMA»), де PUMA означає «Public Use Microdata Area» («Область мікроданих публічного користування»). У якості параметризуючих значень візьмемо коди 12 областей мікроданих – кожне десяте значення в діапазоні 25010 – 25120.

У якості сутнісного атрибуту візьмемо атрибут «Military Service» («Перебування на військовій службі»), значення якого «1» («Активна служба») покладемо сутнісним.

Кількісний сигнал за вказаними вхідними даними представлено на рис. 1 суцільною лінією. Застосувавши один із відомих методів забезпечення групової анонімності [10], отримаємо новий кількісний сигнал, зображений на рис. 1 штрихованою лінією.

Аналізуючи різницевий сигнал (рис. 2), можна зробити висновок, що для забезпечення групової анонімності в мікрофайлі потрібно обміняти параметризуючі значення в $Q = 92$ парах записів, а сам початковий мікрофайл $M$ при цьому розбивається на 7 додатних ($I^+ = \{1,3,4,5,6,8,11\}$) та 5 від’ємних ($I^- = \{2,7,9,10,12\}$) підмікрофайлів. Загальна кількість записів в усіх підмікрофайлах дорівнює 141 838.

Із метою мінімізації втрати корисності мікрофайлу в процесі переміщення вказаного числа записів між підмікрофайлами, у якості визначальних атрибутів оберемо наступні: «Sex» («Стать»), «Age» («Вік»), «Hispanic or Latino Origin» («Іспанське або латиноамериканське походження»), «Marital Status» («Сімейний стан»), «Educational Attainment» («Освіта»), «Citizenship Status» («Громадянство») та «Person's Total Income in 1999» («Сукупний дохід особи з 1999 р.»). Для спрощення інтерпретації результатів модифікації мікрофайлу вважатимемо всі визначальні атрибути номінальними, $\gamma_k = 1 \quad \forall k = \overline{1,7}$, а $\chi_1 = 1$ та $\chi_2 = 0$. У такому разі визначальна метрика (1) показує загальну кількість значень атрибутів респондентів, які потрібно змінити для виконання групової анонімізації.

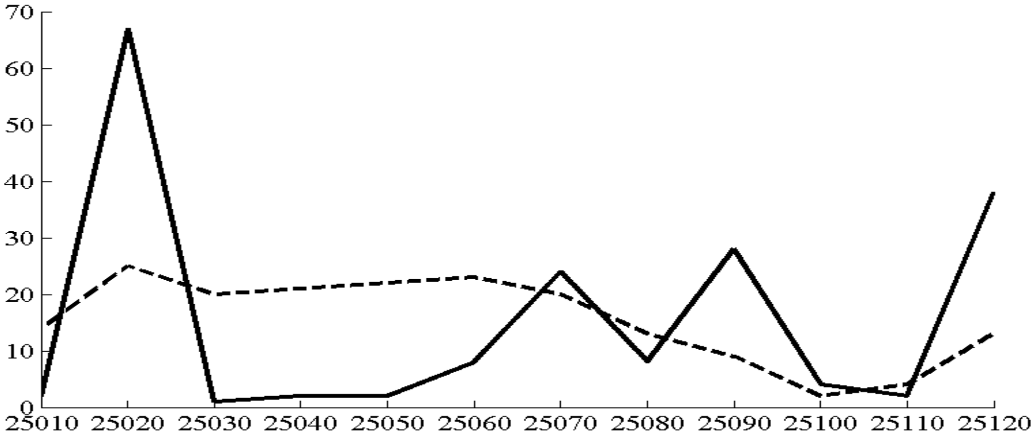



Рисунок 1 – Кількісний (суцільна лінія) та модифікований кількісний (штрихована лінія) сигнали для задачі приховання територіального розподілу військових штату Массачусетс, США

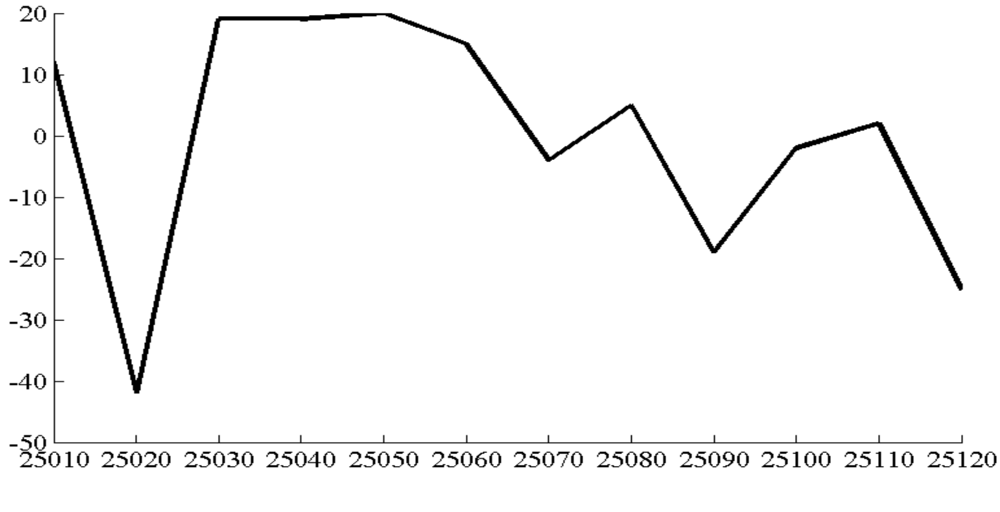



Рисунок 2 – Різницевий сигнал для задачі приховання територіального розподілу військових штату Массачусетс, США

Результати застосування кожної з 18 евристичних стратегій модифікації мікрофайлу представлено в табл. 1.

Оскільки $\mathrm{InfM}_{\max} = 7$, функція пристосованості (2) для меметичного алгоритму з урахуванням (3) набуває вигляду:

$$f(U) = 644 - \sum_{i=1}^{92}\sum_{k=1}^{7}\left(M_{u_{i1}}\left(u_{i2}, J_k\right) \equiv M_{u_{i3}}\left(u_{i4}, J_k\right)\right),$$

де $\mathbf{M}(i,j)$ – оператор повернення значення $j$-о атрибуту $i$-о запису мікрофайлу $\mathbf{M}$, $J_k$ – $k$-й визначальний атрибут.

У якості оператора $R$ було взято оператор порядкового кросовера (order crossover [12]), модифікований для роботи з мультимножинами:

1. Випадковим чином вибрати два рядки $k_1$ та $k_2$ першої батьківської матриці $U_{p1}$.

2. Скопіювати елементи $u_{ij}$, $i = \overline{k_1, k_2}$, $j = \overline{1,4}$ з $U_{p1}$ в $U_{o1}$ на аналогічні позиції.

3. Починаючи з рядка $k_2 + 1$, послідовно знаходити елементи $u_{i1} \in U_{p2}$, які ще можна записати в перший стовпець нащадка $U_{o1}$, не порушуючи структурного обмеження на $U_{o1}$ № 1, та записувати їх разом із відповідним елементом $u_{i2} \in U_{p2}$ у вільні рядки нащадка $U_{o1}$. Вважати, що за останнім рядком матриці $U_{p1}$ ($U_{o1}$) слідує перший.

4. Виконати дії з кроку 3 для елементів $u_{i3}, u_{i4} \in U_{p2}$. При цьому не повинно порушуватися структурне обмеження на $U_{o1}$ № 2.

Нащадок $U_{o2}$ отримується аналогічно, шляхом заміни в вищенаведеному алгоритмі $U_{p1}$ на $U_{p2}$ та навпаки, а $U_{o1}$ – на $U_{o2}$.

У якості операторів $M_1$ та $M_2$ було взято мутацію обміну (swap mutation, [13]), а в якості $M_3$ та $M_4$ – мутацію випадкової заміни (random resetting mutation, [5, c. 43]).

Для відбору батьківських індивідів, а також індивідів, які буде перенесено в популяцію наступного покоління, було застосовано турнірний відбір (tournament selection [14]) із чисельністю турніру 5.

Меметичний алгоритм завершував роботу по генерації 1500 поколінь. Інші параметри меметичного алгоритму було задано наступним чином: $p_c = 1$, $p_{m_1} = p_{m_2} = p_{m_3} = p_{m_4} = 0{,}005$, $p_{mem} = 0{,}75$, $\mu = 100$, $\lambda = 40$.

Для попередження зменшення варіації всередині популяції значення $p_{m_1}, p_{m_2}$, $p_{m_3}, p_{m_4}$ збільшувалися вдесятеро щоразу, коли середньоквадратичне відхилення значень пристосованості індивідів у популяції деякого покоління було менше 1.

Окрім того, якщо в деякому поколінні всі індивіди мали однакову пристосованість, 90% індивідів замінювалися на випадково згенеровані.

Результати застосування меметичного алгоритму для модифікації мікрофайлу представлено в табл. 2. Підрахунки проводилися на комп’ютері з двоядровим процесором Pentium Dual-Core з частотою кожного ядра 2,8 ГГц, а також обсягом оперативної пам’яті 4 ГБ.

Варто зазначити, що меметичний алгоритм в 10 запусках із 50 досяг метрики, меншої за найкращу з тих, які забезпечують евристичні стратегії.

Таблиця 2 – Середні показники застосування евристичних стратегій та меметичного алгоритму до задачі приховання територіального розподілу військових штату Массачусетс, США, за результатами виконання 50 запусків кожного підходу

| № з/п | Назва застосованого підходу | Мінімальне значення визначальної метрики | Середнє значення визначальної метрики | Максимальне значення визначальної метрики | Середній час виконання, с |
|---|---|---|---|---|---|
| 1 | Евристична стратегія 1 | 95 | 103,86 | 111 | 1,282 |
| 2 | Евристична стратегія 2 | 96 | 103,60 | 112 | 1,287 |
| 3 | Евристична стратегія 3 | 96 | 101,96 | 110 | 1,305 |
| 4 | Евристична стратегія 4 | 93 | 100,38 | 107 | 1,335 |
| 5 | Евристична стратегія 5 | 95 | 102,26 | 112 | 1,260 |
| 6 | Евристична стратегія 6 | 96 | 103,68 | 111 | 1,083 |
| 7 | Евристична стратегія 7 | 75 | 81,30 | 88 | 2,690 |
| 8 | Евристична стратегія 8 | 75 | 81,92 | 90 | 2,554 |
| 9 | Евристична стратегія 9 | 73 | 80,70 | 88 | 2,721 |
| 10 | Евристична стратегія 11 | 77 | 77,00 | 77 | 13,998 |
| 11 | Евристична стратегія 12 | 69 | 69,00 | 69 | 13,765 |
| 12 | Евристична стратегія 13 | 62 | 62,00 | 62 | 13,792 |
| 13 | Евристична стратегія 14 | 77 | 77,00 | 77 | 13,029 |
| 14 | Евристична стратегія 15 | 68 | 68,00 | 68 | 13,174 |
| 15 | Евристична стратегія 16 | 75 | 75,00 | 75 | 11,736 |
| 16 | Евристична стратегія 17 | 63 | 63,00 | 63 | 66,811 |
| 17 | Евристична стратегія 18 | 60 | 60,00 | 60 | 71,607 |
| 18 | Евристична стратегія 19 | 59 | 59,00 | 59 | 55,664 |
| 19 | Меметичний алгоритм | 57 | 59,84 | 62 | 2479,022 |

# Висновки

У статті наведено загальну структуру новітнього меметичного алгоритму для модифікації мікрофайлу з мінімізацією спотворень у процесі забезпечення групової анонімності, а також запропоновано низку рекомендацій по підбору його параметрів. На основі порівняння результатів роботи алгоритму та результатів застосування відомих евристичних стратегій до мінімізації спотворень мікрофайлу при розв'язанні задачі приховання територіального розподілу військових штатом Массачусетс, США, можна зробити висновок, що меметичний алгоритм дає можливість отримати кращі результати, але за значно довший проміжок часу, ніж евристичні стратегії. Перспективними є дослідження в напрямку пришвидшення роботи алгоритму за рахунок паралелізації, а також аналіз ефективності алгоритму в залежності від способів реалізації його складових частин.

# Література

1. IPUMS: Minnesota Population Center. Integrated Public Use Microdata Series International [Електронний ресурс]. – Режим доступу : https://international.ipums.org/international/.
2. A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34 [Електронний ресурс] / A. Pfitzmann, M. Hansen. – 2009. – Режим доступу : http://dud.inf.tu-dresden.de/Anon_-Terminology.shtml.
3. Chertov O. Statistical Disclosure Control Methods for Microdata / O. Chertov, A. Pilipyuk // Intern. Symposium on Computing, Communication and Control. – Singapore : IACSIT, 2009. – P. 338-342.
4. Чертов О.Р. Мінімізація спотворень при формуванні мікрофайлу з замаскованими даними / О.Р. Чертов // Вісник Східноукраїнського національного університету імені Володимира Даля. – 2012. – № 8(179). – С. 256-262.
5. Eiben A. E. Introduction to Evolutionary Computing / A.E. Eiben, J.E. Smith. – Springer, 2007. – 316 p.
6. Moscato P. On evolution, search, optimization, genetic algorithms and martial arts: Toward memetic algorithms / Pablo Moscato // C3P Report 826 : Caltech Concurrent Computation Program. – Caltech, CA, 1989. – P. 33-48.
7. Dawkins R. The Selfish Gene / Richard Dawkins. – [3rd ed.]. – Oxford, New York : Oxford University Press, 2006. – 360 p.
8. Chertov O. Providing Group Anonymity in a Microfile with Linguistic Data / O. Chertov, D. Tavrov // Інформаційна безпека. – 2012. – № 2 (8). – P. 168-180.
9. Chertov O. Data Group Anonymity: General Approach / O. Chertov, D. Tavrov // International Journal of Computer Science and Information Security. – 2010. – Vol. 8(7). – P. 1-8.
10. U.S. Census 2000. 5-Percent Public Use Microdata Sample Files [Електронний ресурс]. – Режим доступу : http://www.census.gov/census2000/PUMS5.html.
11. Chertov O. Data Group Anonymity in Microfiles / O. Chertov, D. Tavrov // Вісник інженерної академії України. – 2010. – № 2. – С. 159-164.
12. Syswerda G. Schedule optimization using genetic algorithms / G. Syswerda // Handbook of Genetic Algorithms [ed. L. Davis]. – New York : Van Nostrand Reinhold, 1991. – P. 332-349.
13. Davis L. Applying Adaptive Algorithms to Epistatic Domains / L. Davis // Proceedings of the Ninth International Joint Conference on Artificial Intelligence, 18 – 23 August 1985, Los Angeles, California. : proceedings / [ed. A. Joshi]. – Los Alamos, California : Morgan Kaufmann Publishers, Inc. – 1985. – Vol. 1. – P. 162-164.
14. Brindle A. Genetic algorithms for function optimization : [doctoral dissertation and technical report TR81-2] / Brindle A. – Edmonton : University of Alberta, Department of Computer Science, 1981. – 93 p.

# Literatura

1. IPUMS : Minnesota Population Center. Integrated Public Use Microdata Series International [Elektronnyi resurs]. – Rezhym dostupu: https://international.ipums.org/international/.
2. A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34 [Elektronnyi resurs] / A. Pfitzmann, M. Hansen. – 2009. – Rezhym dostupu : http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.
3. Chertov O. Statistical Disclosure Control Methods for Microdata / O. Chertov, A. Pilipyuk // Intern. Symposium on Computing, Communication and Control. – Singapore : IACSIT, 2009. – P. 338-342.

4. Chertov O.R. Minimizatsiia spotvoren pry formuvanni mikrofailu z zamaskovanymy danymy / O.R. Chertov // Visnyk Skhidnoukrainskoho natsionalnoho universytetu imeni Volodymyra Dalia. – 2012. – № 8 (179). – S. 256-262.
5. Eiben A.E. Introduction to Evolutionary Computing / A.E. Eiben, J.E. Smith. – Springer, 2007. – 316 p.
6. Moscato P. On evolution, search, optimization, genetic algorithms and martial arts: Toward memetic algorithms / Pablo Moscato // C3P Report 826: Caltech Concurrent Computation Program. – Caltech, CA, 1989. – P. 33-48.
7. Dawkins R. The Selfish Gene / Richard Dawkins. – [3rd ed.]. – Oxford, New York : Oxford University Press, 2006. – 360 p.
8. Chertov O. Providing Group Anonymity in a Microfile with Linguistic Data / O. Chertov, D. Tavrov // Informatsiina bezpeka. – 2012. – № 2 (8). – P. 168-180.
9. Chertov O. Data Group Anonymity: General Approach / O. Chertov, D. Tavrov // International Journal of Computer Science and Information Security. – 2010. – Vol. 8 (7). – P. 1-8.
10. U.S. Census 2000. 5-Percent Public Use Microdata Sample Files [Elektronnyi resurs]. – Rezhym dostupu : http://www.census.gov/census2000/PUMS5.html.
11. Chertov O. Data Group Anonymity in Microfiles / O. Chertov, D. Tavrov // Visnyk inzhenernoi akademii Ukrainy, 2010. – № 2. – S. 159-164.
12. Syswerda G. Schedule optimization using genetic algorithms / G. Syswerda // Handbook of Genetic Algorithms [ed. L. Davis]. – New York : Van Nostrand Reinhold, 1991. – P. 332-349.
13. Davis L. Applying Adaptive Algorithms to Epistatic Domains / L. Davis // Proceedings of the Ninth International Joint Conference on Artificial Intelligence, 18 – 23 August 1985, Los Angeles, California. Volume 1 : proceedings / ed. A. Joshi. – Los Alamos, California : Morgan Kaufmann Publishers, Inc., 1985. – P. 162-164.
14. Brindle A. Genetic algorithms for function optimization : [doctoral dissertation and technical report TR81-2] / A. Brindle. Edmonton : University of Alberta, Department of Computer Science, 1981. – 93 p.

***O.R. Chertov, D.Y. Tavrov***

## *Memetic Algorithm for Microfile Modification With Minimizing Distortion while Providing Group Anonymity*

Modern information technologies enable analyzing huge amounts of primary non-aggregated digital data, which can be used to generate hypotheses about data distribution and validate them. The IPUMS-International project, which contains 480 million records about respondents of 211 censuses from 68 countries, is a prominent example of the primary data accessibility. Publishing such data increases the threat of disclosing sensitive information.

To protect information on a single person, one needs to provide individual data anonymity, whereas providing group data anonymity presupposes protecting intrinsic data properties and distributions. Group anonymity methods need not only protect the underlying data distribution, but also ensure sufficient data utility after their transformation. The latter task belongs to the class of exhaustive search problems, and may be solved by utilizing heuristics procedures.

Evolutionary algorithms are heuristic guided random search methods mimicking biological evolution by natural selection. To perform successfully, they have to be inherently stochastic, which turns out to be a downside when converging to an optimum. Memetic algorithms mimic cultural evolution, and are a combination of evolutionary algorithms and local search procedures. Applying local search enables convergence to the optimum and enhances overall algorithm performance by incorporating problem specific knowledge.

In the paper, we propose a novel memetic algorithm for modifying statistical data micro-files when providing group anonymity. Its application to a real data based problem of protecting military personnel regional distribution for the state of Massachusetts (the US) exhibits better results in terms of solution quality compared to the previously applied heuristics, however, at the cost of longer time needed to run the algorithm.

*Стаття надійшла до редакції 05.04.2013.*